基于粒划分方法构建决策树的算法研究

刘 军
(南京工业大学 电子与信息工程学院,江苏 南京 210009)

摘 要:针对当前基于信息增益和粗集属性约简作为属性选择标准建树算法存在的不足,以粒划分方法为理论基础,将属性按其取值划分为若干属性粒,提出以属性粒的长度量和其所对应决策属性的粒类别个数作为确定分裂属性的基本参数,自顶向下逐级构造决策树,不涉及信息增益、等价类和属性约简等复杂运算的中间过程。该算法的优点在于不仅考虑本层结点的划分而且预测下层结点的走向,具有较高的精准度,而且解决了当前建树算法不具有普遍适应的难题。

关键词:粗糙集;决策树;粒划分
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2012)10-0087-04

Research on Algorithm of Constructing Decision Tree Based on Granulatio Division Method

LIU Jun
(College of Electronics and Information Engineering of Nanjing University of Technology, Nanjing 210009, China)

Abstract: In view of the current algorithm building decision tree has shortcomings by attribute as selection standard based on information gain and attribute reduction of rough set, the attribute according to the attribute value is divided into a number of granulation by the granulation division method as theoretical basis, put forward attribute granulation length and the corresponding decision attribute granulation class number as basic parameters of determined splitting attribute and stepwise construct decision tree from top to down. The algorithm does not involve complex operation process with the information gain, equivalence classes and attributes reduction. The advantage of this algorithm is that not only considers the layer nodes division but also predicts the trend of lower nodes, has high precision and solves the problem that the current algorithm of building tree is not universal adaptation.

Key words: rough set; decision tree; granulatio division

1 概 述

决策树采用树形结构来对决策表进行分类,决策树的建树算法具有思想简单和识别样本效率高的特点,是一种简洁且易理解的智能分类表示方法。但当前的决策树算法还存在以下问题:

(1)经典的ID3算法[1]用信息增益作为属性选择的标准构造的决策树平均深度较小,分类速度较快,但这种算法偏向于选择属性值较多的属性,是一种精度不高的策略[2]。

(2)基于粗集属性约简的决策树,由于建树前进行属性约简[3,4],属性约简是一种粒度较大的约简,往往会约简了对建树重要的属性。实质上,属性是由各种属性值组成,属性值的大部分不重要并不等于其每一部分都不重要。另外约简属性的算法仍存在许多难题[5]。

针对上述问题,文中以粒[6]划分为理论基准,将属性 Ai 分成若干基本粒,以基本粒对 D 的划分能力为依据确定分裂属性,以分裂属性逐级划分建树,使可尽快分出叶结点的粒优先,所建树上层叶结点多且深度减少。

2 理论和实现

2.1 理 论

设决策表 $S=(U,An,D,V,F)$[7],子集 An 和 D 为条件属性和决策属性;$U=\{u1,u2,\cdots,um\}$ 是对象的集合;V 是属性值集;F 是一个信息函数。

定义1[6] (Ai,v) 称为决策表 S 中的属性基本粒;而 (D,d) 称为 S 中的决策属性基本粒。

一般条件属性 Ai 和决策属性 D 由多个基本粒组成,而 Ai 的各基本粒集对应 D 的各基本粒集的关系是确定 Ai 划分能力[8] 及最终构建决策树的理论依据。

收稿日期:2012-02-03;**修回日期**:2012-05-11
基金项目:国家自然科学基金资助项目(60673185);教育部留学回国人员科研启动基金资助项目(教外司留[200711108号])
作者简介:刘 军(1977-),女,讲师,硕士,研究方向为数据挖掘、粗糙集、粒计算。

定义 2 $LAiv = \mathrm{cnt}(ui \in U \mid (Ai,v))$ 为属性 Ai 中 $V = v$ 时属性值个数,称为 (Ai,v) 的粒长度[9]。

$LAiv$ 是个独立的参数,Ai 中的每种取值 v 都有一个 $LAiv$ 与之对应。

定义 3 $DAiv = \mathrm{Ccount}(xi \in U \mid (Ai,v), d \in D))$ 是 (Ai,v) 对应与 $d \in D$ 的类别个数,称为 (Ai,v) 的粒别值。

$DAiv$ 表示 (Ai,v) 可划分 D 的强度,是 Ai 和 D 之间的关联参数。

2.2 实 现

运用粒划分方法构建决策树首要问题是确定分裂属性,即如何运用 $LAiv$ 和 $DAiv$ 在 An 中选取最优属性[10],属性 $A_i \in \{A_n\}$ 是由若干 (Ai,v) 组成,所以先确定各 (Ai,v) 的划分能力,再由各 (Ai,v) 确定属性 Ai 的划分能力。

从建树资源上分析,可将 $LAiv$ 视为可分配量,而 $DAiv$ 视为受配类别量。$LAiv$ 值越大且 $DAiv$ 值越小,则 (Ai,v) 的划分能力越大。所以 $LAiv$ 和 $DAiv$ 的值确定 (Ai,v) 划分能力:若 $DAiv = 1$,则结点划分后,可一次划分出 $LAiv$ 行,即得 1 枝 1 叶;若 $DAiv = 2$,首次不能划分出叶结点,增加 1 枝后,下一次枝结点划分后,可划分出 2 枝 2 叶。以此类推,得 (Ai,v) 划分力 $LDAiv$ 算式:

$$LDAiv = LAiv/(DAiv + L) = LAiv/(2DAiv - 1) = LAiv/(DiAiv)$$

式中,L 是建树的层次系数:$L = (DAiv - 1)$。

各 (Ai,v) 划分能力之和即为属性 Ai 的划分能力 GAi:

$$GAi = \sum LDAiv = \sum (LAiv/DiAiv)$$

所以取 GAi 最大者即为分裂属性:

$$Aj = \mathrm{Max}(GAi) \{(i = 1,2,\cdots,n) \mid j \in n\}$$

依据上述算法,逐层进行分裂属性 Aj 选择,直至 Aj 中只有一种 d 值,则生成最终叶结点:

```
Create Dt (An,D,GA){// 程序入口
if(cnt(D) = cnt(GA)){ // 判 D 是否仅一类 d 值
Create Nd(GA,D);return}// 建立一叶结点
for( ;Ai;){// 对{An} 中的每个属性 Ai 循环
for( ;v ∈ Ai;)// 对 Ai 中的每种(Ai,v) 循环
GAi += (LAiv/DiAiv);// 求属性 GAi 的划分度
if(GAi > {GAj}) Aj = Ai;}// 求分裂属性 Aj
CreateNd(Aj);// 用 Aj 构建 1 个划分结点
for( ;v ∈ Aj;){// 对 Aj 中的每种(Ai,v) 循环
U = {xi} | (v ∈ Aj);// 求 Aj 中等于 v 值的行号
{An,D} | U = {xi};// 按 U 对{An,D} 划分
{An} -= Aj;// 在{An} 去除属性 Aj
Create Dt(An,D,GAj);}}// 调用 Create Dt。
```

3 实例比较分析

3.1 与基于粒度商算法的比较分析

文献[11]中表 2 是一个决策系统,见表 1。

●首次分裂属性 Aj 的选择:

$$Ga1 = \sum Ga1v \mid (v \in a1)$$

$$= \sum (La1v/Dia1v) \mid (v \in a1)$$

$$= \sum Ga1v(v = 1,2,3,4,5)$$

$$= (La11/Dia11) + (La12/Dia12) + (La13/Dia13) + (La14/Dia14) + (La15/Dia15)$$

$$= 4/(2+1) + 6/(3+2) + 6/(2+1) + 7/(3+2) + 3/(2+1) = 6.93;$$

$$Ga2 = 6/(3+2) + 6/(3+2) + 6/1 + 5/(2+1) + 3/(2+1) = 11.06;$$

$$Ga3 = 6/(3+2) + 7/(3+2) + 7/(2+1) + 3/1 + 3/(2+1) = 8.93;$$

$$Ga4 = 10/(2+1) + 3/(2+1) + 3/(2+1) + 4/1 + 6/(3+2) = 10.53。$$

$a2 = \max(Gai)$ $(i = 1,2,3,4)$。

经 a2 划分后,第 0 层得 1 个叶结点:a2 = 3 d = 0。

●在{a1,a3,a4} | (a2 = 1)集中,分裂属性 Aj 的选择:

$$Ga1 = 1/1 + 5/(3+2) = 2;$$

$$Ga3 = 1/1 + 3/(2+1) + 2/1 = 4;$$

$$Ga4 = 1/1 + 1/1 + 2/1 + 2/(2+1) = 5.3。$$

$a4 = \max(Ga\,i)$ $(i = 1,3,4)$。

经 a4 划分后,得在 a2 = 1 的约束下第 1 层的 3 个叶结点:a4 = 3 d = 2;a4 = 4 d = 1; a4 = 5 d = 2。

●在{a1,a3,a4} | (a2 = 1) ∩ (a4 = 1)集中,分裂属性 Aj 的选择:

$$Ga1 = 2/(2+1) = 0.7;\ Ga3 = 1/1 + 1/1 = 0.2。$$

$a3 = \max(Ga\,i)$ $(i = 1,3)$。

经 a3 划分后,得在(a2 = 1) ∩ (a4 = 1)的约束下第 2 层的 2 个叶结点:a3 = 2 d = 0;a3 = 3 d = 1。

以上是所建树的一个分枝,类似地,在{a1,a3,a4} | (a2 = 2)集中,分裂属性 Aj 是 a3(第 1 层)、a4(第 2 层);在{a1,a3,a4} | (a2 = 3)集中,分裂属性 Aj 是 a2(第 1 层);在{a1,a3,a4} | (a2 = 4)集中,分裂属性 Aj 是 a4(第 1 层);在{a1,a3,a4} | (a2 = 5)集中,分裂属性 Aj 是 a1(第 1 层)。

表示为决策树(以磁盘文件树格式表示)如下:

```
<a2>
a2.1  <a4>a4.1  <a3>:a3.2  d.0;a3.3  d.1
                a4.3  d.2;a4.4  d.1;a4.5  d.2
a2.2  <a3>a3.1  <a4>:a4.1  d.0;a4.5  d.2
```

a3.2　d.0;a3.3　d.0;a5.5　d.1

a2.3　d.0

a2.4　<a4>a4.2　d.2;a4.4　d.1;a4.5　d.1

a2.5　<a2>a2.1　d.1;a2.3　d.0

文献[11]中图 2 所示。将 a4 定为首选划分属性所建的树不精准,比文中算法首选 a2 所建树多了 2 个树枝。主要原因是文献[11]判定式中(如 QDa1　d=|a1|/|a1∪D|)只考虑本层单个属性 ai 对决策属性 D 的划分强度,而没有考虑(预测)后续层的划分强度。

表 1　决策表

U	a1	a2	a3	a4	d	U	a1	a2	a3	a4	d	U	a1	a2	a3	a4	d
1	1	2	1	1	0	10	3	2	2	1	0	19	5	3	5	1	0
2	1	3	5	5	0	11	2	2	3	1	0	20	5	2	5	1	1
3	1	4	1	5	1	12	4	1	1	4	1	21	3	3	3	3	0
4	1	5	1	5	1	13	4	1	2	1	0	22	3	1	3	4	1
5	2	2	1	1	0	14	4	4	2	2	2	23	3	4	3	5	1
6	2	2	1	5	2	15	4	1	2	3	2	24	5	3	4	1	0
7	2	3	3	2	0	16	4	4	2	4	1	25	3	5	4	2	0
8	2	3	4	1	0	17	4	1	2	5	2	26	3	5	2	3	0
9	2	4	3	4	1	18	4	1	3	1	1						

3.2　与属性加权分类粗糙度的算法的比较分析

对文献[2]的表 1 计算首选分裂的过程如下:

$Ga1=\sum Ga1v\mid (v\in a1)$

$=\sum(La1v/Dia1v)\mid(v\in a1)=\sum Ga1v(v=1,2)$

$=(La11/Dia11)+(La12/Dia12)$

$=8/(2+1)+4/(2+1)=4$;

$Ga2=3/(2+1)+6/(2+1)+3/(2+1)=4$;

$Ga3=1/1+8/(2+1)+3/(2+1)=4.66$;

$Ga4=4/(2+1)+6/(2+1)+2/1=5.33$。

$a4=\max(Ga\,i)\ (i=1,2,3,4)$。

对本例而言,与文献[2]的算法结论相同但本算法简洁。

3.3　与正区域算法的比较分析

按本算法(二级判定)对文献[12]的表 1 计算如下:

$Ga1=3/1+2/(1+1)+2/(2+2)+2/1+3/(2+2)+1/(1+1)=7.75$;

$Ga2=5/1+1(1+1)+2/(1+1)++2/(1+1)+1/(1+1)+1/(1+2)+1/(1+2)=8.66$;

$Ga3=2/1+3/(1+1)+2/(2+2)+2/(2+1)+3/(1+1)+1/1=7.16$;

$Ga4=3/(1+1)+2/(1+1)+1/(1+1)+3/1+4/1=10$。

$a4=\max(Ga\,i)(i=1,2,3,4)$。

由 a4 作为首选分裂属性建树如下:

<a4>

a4.0　<a2>a2.1　D.0

a2.2　D.1

a2.3　D.2

a4.1　D.0

a4.2　<a3>a3.2　D.1

a3.3　D.2

该树(首选 a4)2 层 8 个树枝优于文献[12]中图 1 树(首选 a2)的 2 层 10 个分枝,而首选 a1 和 a3 都是 3 层 12 个分枝。这个结论符合上述 Gai 的判定结论。

3.4　与基于粒度的粗集算法的比较分析

将文献[13]中表 3 按属性分排如文中表 2。

表 2　决策表及各属性分列表

U	a	b	c	d	e	U	a	e	U	b	e	U	c	e	U	d	e
1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	8	1	3
2	1	2	1	2	1	2	1	1	10	1	3	2	1	1	9	1	3
3	2	2	1	2	2	3	2	2	2	2	1	3	1	2	10	1	3
4	2	3	3	2	2	4	2	2	3	2	2	5	2	2	2	2	1
5	2	2	2	3	2	5	2	2	5	2	2	6	2	2	3	2	2
6	3	2	2	2	2	9	2	3	6	2	2	8	2	3	4	2	2
7	3	2	3	2	2	6	3	2	7	2	2	9	2	3	6	2	2
8	3	3	2	1	3	7	3	2	4	3	2	10	2	3	7	2	2
9	2	3	2	1	3	8	3	3	8	3	3	4	3	2	1	3	1
10	3	1	2	1	3	10	3	3	9	3	3	7	3	2	5	3	2

按表 2 的排序计算如下:

$Ga=\sum Gav\mid(v\in a)=\sum(Lav/Diav)\mid(v\in a)$

$=\sum Gav(v=1,2,3)$

$=(La1/Dia1)+(La2/Dia2)+(La3/Dia3)$

$=2/1+4/(2+1)+4/(2+1)=4.67$;

$Gb=2/(2+1)+5/(2+1)+3/(2+1)=3.33$;

$Gc=3/(2+1)+5/(2+1)+2/1=4.67$;

$Gd=3/1+5/(2+1)+2/(2+1)=5.33$。

$d=\max(Gx)\ (x=a,b,c,d)$

由 d 作为首选分裂属性建树如下:

<d>

d.1　e.3

d.2　<a>a.1　e.1

a.2　e.2

a.3　e.2

d.3　<a>a.1　e.1

a.2　e.2

该树(首选 d)2 层 8 个树枝优于文献[13]中图 1 所示树(首选 a)的 2 层 11 个分枝。文献[13]先对属性集约简得最小约简{a,b,c},之后在{a,b,c}选优属性建树。首先,因为{a,d}完全可划分出 e,所以文献[13]中表 3 的最小约简是{a,d}而不是{a,b,c};其次它在{a,b,c}集建树不能用属性 d(因为被约简)从而使所建树不精准。

3.5 与最简规则比较分析

表 2 的最简规则(表 3 左)和其决策树(见 3.4)展开规则(表 3 右)相比较,两者都只涉及{a,d}属性;展开规则个数多 1 个。若排除树规则有重值的因素,决策树展开规则与最简规则基本接近,可见决策树是较优的决策树。

表 3 最简规则与树展开规则比较

d. 1->e. 3	d. 1->e. 3
d. 2∧a. 2->e. 2	d. 2∧a. 1->e. 1
d. 2∧a. 3->e. 2	d. 2∧a. 2->e. 2
	d. 2∧a. 3->e. 2
d. 3∧a. 2->e. 2	d. 3∧a. 1->e. 1
a. 1->e. 1	d. 3∧a. 2->e. 2

4 结束语

(1)粗糙集是以 $\{An\}$ 与 D 划分关系为出发点,而本算法是以 $Ai \in \{An\}$ D 划分关系为出发点,首要解决的是在 $\{An\}$ 中确定分裂属性而并非确定核属性。

(2)将 Ai 按属性值划分为若干粒,提出以粒为基本单位确定分裂属性且直接建树,不涉及等价类和属性约简等复杂运算的中间过程。

(3)算法具有可扩展性,当属性划分力接近时,可采用扩级方法进一步确定分裂属性。

参考文献:

[1] 杨 静. 决策树算法的研究与应用[J]. 计算机技术与发展,2010,20(2):114-120.

[2] 丁春荣. 基于粗糙集的决策树构造算法[J]. 计算机工程,2010,36(11):75-77.

[3] Wang X J, Reyes J L, Chua Nam-Hai, et al. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNAtargets[J]. Genome Biol. ,2004(5):1-15.

[4] 覃伟荣. 基于粗糙集理论的条件属性动态约简算法[J]. 计算机技术与发展,2008,18(8):23-25.

[5] 李永华. 一种基于 rough 集的属性约简的改进算法[J]. 计算机应用,2008,28(8):200-202.

[6] 李 鸿. 知识粗糙性与知识粒度的关系研究[J]. 计算机技术与发展,2007,17(8):117-119.

[7] Kim Sung-Kyu, Nam Jin-Wu, Rhee Je-Keun, et al. miTarget: microRNA target gene prediction using a support vector machine[J]. Bioinformatics,2006(7):411-422.

[8] 葛 浩. 一种改进的基于二进制可分辨矩阵属性约简算法[J]. 计算机技术与发展,2008,18(8):12-15.

[9] 王国胤,张清华,胡 军. 粒计算研究综述[J]. 智能系统学报,2007,2(6):8-26.

[10] Chou Chi-Hung, Yang Tsung-Hsien, Tsao Shih-Chiang, et al. Standard operating procedures for embedded linux systems [J]. Linux Journal,2007(160):10-14.

[11] 周 军. 基于粒度商的决策树构造算法[J]. 计算机工程与设计,2009,30(16):3826-3829.

[12] 高 静. 一种改进的基于正区域的决策树算法[J]. 计算机科学,2008,35(5):138-142.

[13] 陈 婷. 基于粒度的粗集-决策树雷达信号识别模型[J]. 微电子学与计算机,2008,25(12):13-16.

(上接第 86 页)

下一步主要工作是对任何笔画都能较好地处理,主要思想:

(1)不能处理的复杂笔画都是由简单曲线笔画类型或组合曲线笔画类型拼合而成,因此对于不能处理的复杂曲线笔画类型把它分解成两个或更多的训练集库中简单曲线笔画类型或组合曲线笔画类型来分段对其处理。

(2)对一些奇怪的曲线笔画类型单独做成美化曲线笔画类型库,当遇到不能处理的笔画优先去美化曲线笔画类型库中查找处理。

参考文献:

[1] 林 民,宋 柔. 一种笔段网格汉字字形描述方法[J]. 计算机研究与发展,2010,47(2):318-327.

[2] Ideographic Description[EB/OL]. 2003. http://www.unicode.org/versions/Unicode4.0.0/ch11.pdf.

[3] Cook R. A Specification for CDL(Character Description Language):an extract of[D]. USA:UC Berkeley, Dept. of Linguistics,2003.

[4] 傅用和. 中文信息处理[M]. 广州:广东教育出版社,1999.

[5] 林 民,宋 柔. 汉字字形形式化描述方法及应用研究[D]. 北京:北京工业大学,2009.

[6] 罗笑南,王岩梅. 计算机图形学[M]. 广州:中山大学出版社,2003:188-196.

[7] 林 民,宋 柔. 汉字的笔段网格字形描述及字形比对计算[J]. 计算机辅助设计与图形学学报,2009,21(9):1298-1307.

[8] 林 民,宋 柔. 一种面向构形计算的汉字字形形似化描述方法[J]. 中文信息学报,2008,22(3):115-123.

[9] Zhang Maiku, Lin Min, Huang Hanquan. Beautification of Chinese Character Stroke-Segment-Mesh Glyph Stroke Curve [J]. Advances in MSEC,2012,2:101-110.

[10] 上海交通大学汉字编码组. 汉字信息字典[M]. 北京:科学出版社,1988.

[11] 国家语言文字工作委员会. GF3001-1997 信息处理用 GB13000.1 字符集汉字部件规范[S]. 北京:语文出版社,1997.

基于粒划分方法构建决策树的算法研究

作者：刘军
作者单位：南京工业大学 电子与信息工程学院,江苏 南京 210009
刊名：计算机技术与发展
英文刊名：Computer Technology and Development
年，卷(期)：2012(10)